Improving Patient Prostate Cancer Risk Assessment: Moving From Static, Globally-Applied to Dynamic, Practice-Specific Cancer Risk Calculators

Andreas N. Strobl, Andrew J. Vickers, Ben van Calster, Ewout Steyerberg, Robin J. Leach, Ian M. Thompson, Donna P. Ankerst

From the Departments of Mathematics [ANS,DPA] of the Technical University Munich and Helmholtz Zentrum, Munich, Germany; the Departments of Cellular and Structural Biology [RJL], Urology[IMT,RJL,DPA], Epidemiology and Biostatistics [DPA] of the University of Texas Health Science Center at San Antonio, San Antonio, Texas, USA; Department of Public Health, Erasmus MC, Rotterdam, Netherlands[ES]; Department of Development and Regeneration, KU Leuven, Belgium [BvC]; Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York City, New York, USA [AJV]

Corresponding author: Andreas Strobl, M.Sc; Department of Mathematics, Technische Universität München, M12, Boltzmannstr. 3, D-85747 Garching near Munich; Email: a.strobl@tum.de; Phone: +49 89 289-18390; Fax: +49 89 289-18435.

Support and financial disclosure declaration: This work was sponsored by grants R01CA179115, U01CA86402, 5P30 CA0541474-19, UM1 CA182883, P01 CA108964-06, the Prostate Cancer Foundation, the Sidney Kimmel Center for Prostate and Urologic Cancers, P50-CA92629, and P30-CA008748. No authors have a financial conflict of interest with this manuscript.

ABSTRACT (193; max 200)

Clinical risk calculators are now widely available but have generally been implemented in a static and one-size-fits-all fashion. The objective of this study was to challenge these notions and show via a case study concerning risk-based screening for prostate cancer how calculators can be dynamically and locally tailored to improve on-site patient accuracy. Yearly data from five international prostate biopsy cohorts (3 in the US, 1 in Austria, 1 in England) were used to compare 6 methods for annual risk prediction: static use of the online US-developed Prostate Cancer Prevention Trial Risk Calculator (PCPTRC); recalibration of the PCPTRC; revision of the PCPTRC; building a new model each year using logistic regression, Bayesian prior-to-posterior updating, or random forests. All methods performed similarly with respect to discrimination, except for random forests, which were worse. All methods except for random forests greatly improved calibration over the static PCPTRC in all cohorts except for Austria, where the PCPTRC had the best calibration followed closely by recalibration. The case study shows that a simple annual recalibration of a general online risk tool for prostate cancer can improve its accuracy with respect to the local patient practice at hand.
Keywords: prediction, discrimination, calibration, prostate cancer, logistic regression, revision
Running title: Dynamic tailored risk prediction
Text (3758, max 4000)

1. INTRODUCTION

Clinical risk prediction tools are now widely available on the internet and provide a valuable decision-aid to doctors and patients regarding treatment choices. There are currently hundreds of clinical risk prediction tools available online, with objectives ranging from the prediction of onset of disease for use in screening to prognosis of outcomes following treatment for disease [1-3]. Interestingly, despite the recent interest in personalized approaches to medicine, the big data daily flowing into clinical practices, and changes in patient populations and clinical practice over time, these risk calculators have generally remained static and applied in a one-size-fits-all fashion. For instance, 2013 US national guidelines for the prevention of cardiovascular diseases prescribed statins for persons with elevated risk based on a global score that was developed using a pooled cohort of patients monitored from the late 1980s to the early 2000s [4]. Subsequent validations on five external cohorts showed that the recommended risk score would greatly overestimate actual risk on contemporary populations, with up to 40 to 50% of the millions classified as high-risk in fact over-prescribed [5]. The widespread availability of electronic medical data raises the possibility that such models could instead evolve over time, automatically changing in tandem with evolving global clinical practice patterns [6]. Within individual hospitals, the ability to capitalize electronic medical record (EMR) data would additionally permit tailoring of a global risk tool to the hospital-specific patient population at hand, for example, allowing a different dynamic evolution of predictions for high-risk clinically referred versus healthy screening institutions.

As the case study to be investigated in this article, the Prostate Cancer Prevention Trial Risk Calculator (PCPTRC) is a static risk tool that predicts the likelihood of detecting prostate cancer if a prostate biopsy were to be performed. It uses as inputs the commonly collected clinical risk factors: prostate-specific antigen (PSA), digital rectal exam (DRE), age, race, family history of prostate cancer, and prior biopsy history [3]. The model it is based on was developed using prostate biopsy data from participants on the placebo arm of a very unique prostate cancer prevention trial, the PCPT [7]. The PCPT provided the only patient population ever to be free of selection bias because at the end of seven years on the study all participants were requested to undergo prostate biopsy even if they lacked a clinical indication for biopsy (n = 5519) [8]. The posting of the calculator online in 2006 facilitated subsequent external validation on a range of cohorts that differed both in terms of patient composition and date of collection [9-21]. The latter was important since a shift in prostate biopsy practice occurred after the PCPT was completed:

the number of sampled tissue cores on biopsy increased from 6 cores (3 on each side) in the PCPT to the now contemporary practice of 12 cores (6 on each side). It has been documented that a greater number of biopsy cores retrieved at biopsy increases the chance of detection of prostate cancer [22].

Statistical approaches to updating an existing risk prediction tool have been proposed, ranging from simple adjustment of the intercept of a model to re-estimation of multiple coefficients in the original model [23]. One-time updating approaches have been implemented in a variety of clinical settings, resulting in improved diagnostic or prognostic performance [24-29]. The need for continual temporal recalibration of a risk tool has been emphasized [30, 31], along with the concept of transfer learning from similar hospitals when sample sizes at individual institutions are low [32].

In an era where patient data are housed electronically, risk prediction tools could and should be automatically updated with local data as soon as such data arrive. The objective of this study was to challenge the ubiquitous notion of static universal risk prediction and show via a case study how prediction can easily be adapted to the patient data on-site, and thus improve the accuracy of prediction for local patients.

2. METHODS

2.1. Participants and biopsy results

Five international cohorts from the Prostate Biopsy Collaborative Group (PBCG) were used to compare various methods for developing an institution-specific risk calculator. These have been previously described [21]. Three screening cohorts, the San Antonio Center of Biomarkers of Risk for Prostate Cancer study (SABOR), Texas, U.S., ProtecT, UK, and Tyrol, Austria followed primarily a 10-core biopsy scheme. Two clinical cohorts from the U.S., Cleveland Clinic, Ohio and the Durham VA, North Carolina, comprised patients referred for clinical symptoms. Those three cohorts used mixed biopsy schemes, but primarily 10- to 14-cores. Not all cohorts had all of the PCPTRC risk factors available; only those risk factors that were missing in less than 15% of the cases were used in the analysis. Biopsy records with associated PSA values higher than 50 ng/ml or with unknown Gleason grade were excluded. If cohorts had only few biopsies in the beginning and ending years, those years were aggregated

into the first and last year. The number of biopsies per year in the resulting data set ranged from 73 (Durham) to 1106 (ProtecT).

2.2. PCPTRC

A modification of version 2.0 of the PCPTRC was used for the methods that tailored an existing risk tool [33]. While PCPTRC 2.0 provides separate estimates of the risks of low- versus high-grade prostate cancer, for this study a logistic regression of any prostate cancer was performed using the same dataset and the same covariates as the PCPTRC model: PSA, age, DRE, first-degree family history of prostate cancer, race (African American versus not) and history of a prior biopsy. When a risk factor was missing in more than 15% of biopsies in a cohort, it was not used in the analysis. This was the case for three of the binary covariates: African American race, prior biopsy and family history. Eight separate logistic regressions were run for each possible combination of missing values from these three variables and the corresponding model used for the cohort. The PCPTRC logistic regression models are given in Table 1 of the Supplementary Appendix.

2.3. Validation sets and metrics

The different statistical methods for annually updating a risk tool were compared using each consecutive year, starting with year 2, as the validation set, and all past years as a training set. In this manner the training set grew cumulatively in size with each year and the validation set changed each year. To compare methods in absence of a fluctuating validation set, the process was repeated using a fixed validation set consisting of the biopsies in the last three years of each cohort. The methods were compared in terms of discrimination and calibration. Discrimination was measured using the area-underneath-the-receiver-operating-characteristic-curve (AUC), which equals the probability that for a randomly chosen cancer case/control pair, the case has a higher predicted risk of cancer. AUCs vary from 50% (chance discrimination) to 100% (perfect discrimination), with higher values indicating better discrimination. Ninety-five percent confidence intervals (95% CI) for AUCs were calculated using non-parametric U-statistics as commonly implemented in statistical packages. Calibration was measured via the Hosmer-Lemeshow statistic (HLS), which provides a single summary of the commonly used calibration plots. For each method of estimating risk, patients in the validation set were grouped into ten decile groups according to estimated risk: patients with the lowest 10^{th} percentile of risks, risks in

the 10th to 20th percentile and so on up to patients with the highest 10th percentiles of risks. The observed rate of prostate cancer in each of the decile groups was computed (O_g) and compared to the mean of the n_g estimated risks in each decile group (E_g). The HLS equals the sum $\sum_{g=1}^{10} \frac{n_g (O_g - E_g)^2}{E_g(1-E_g)}$, with larger values indicating poorer fit; 95% CIs for the HLS were computed using the approximation that the HLS follows a chi-square distribution with degrees of freedom equal to 8.

2.4. Statistical methods

Details of the individual methods follow.

2.4.1. PCPTRC:

This method performed no model building or augmentation and thus tests the value of a static model. For each individual in the training set the PCPTRC score was computed, allowing for missing values for some of the variables; see Supplementary Appendix, Table 1.

2.4.2. Recalibration:

This method performed a logistic regression on the training set using the PCPTRC linear predictor $pred_{PCPTRC,i} = \beta'_{PCPTRC} X_i$ as the only variable. The intercept and slope of the resulting linear predictor $lp_{update,i} = \alpha_{update} + \beta_{update} pred_{PCPTRC,i}$ indicated how well the PCPTRC was calibrated to the training sample. An intercept of 0 and a slope of 1 corresponded to perfect calibration. The risk of prostate cancer in the test set was $1/\{1+\exp(-lp_{update,i})\}$.

2.4.3. Logistic regression:

For this method a new logistic regression model was built using the training data and all PCPTRC risk factors age, race, PSA, DRE, family history, and prior biopsy history that were available in the training set (a "clean-slate" approach).

Revision: In this method, not only the PCPTRC risk factors, but also the linear predictor of the PCPTRC was allowed to enter the logistic regression as a potential variable [23]. Model selection was performed using the stepwise Bayesian Information Criterion (BIC) to arrive at a logistic regression model with linear predictor $\alpha_{update} + \beta'_{update} X_i$, where X_i is a vector of predictors that contains $pred_{PCPTRC,i}$ and any other available PCPTRC risk factors. Stepwise regression was initiated separately with an intercept only model and with the model including all possible

variables. The model with lower BIC was selected for estimation of the risk of prostate cancer in the test set.

2.4.4. Bayesian method:

This approach was based on a logistic regression model assumed for the training data to form the data likelihood, and with a prior $\pi(\beta)$ for the vector of log odds ratios. The set of participants and variables were reduced so that all patients had all variables measured and the models could be fit using Markov Chain Monte Carlo (MCMC), which is implemented as part of the MCMCpack package in the R statistical software. The prior for β for each year was assumed to be multivariate normal. The prior mean was set to be the PCPTRC estimated coefficients. The prior variance matrix was set to be the estimated variance-covariance matrix of log odds ratios from the PCPTRC multiplied by the sample size of PCPTRC to dilute the information and yield a unit-information prior [34].

2.4.5. Random forests:

Random forests are a combination of many "trees", where each regression tree starts with a root node containing the most influential covariate, finds the optimal cut point split on that covariate, and continues splitting subsequent branches by other covariates [35]. Trees are built from random bootstrap samples from the data set. We used the R package randomForest which implements the Breiman algorithm, using the default settings, including 500 trees. All available PCPTRC risk factors were allowed for the building of individual trees. We investigated an option whereby the PCPTRC linear predictor was also allowed, making this method a form of non-parametric revision. However, this turned out not to perform well due to the high correlation between the PCPTRC linear predictor and PSA, and the PCPTRC predictor was subsequently not allowed for inclusion. For prediction of cancer for a new individual, the percent of trees classifying the individual as a cancer case was used.

3. RESULTS

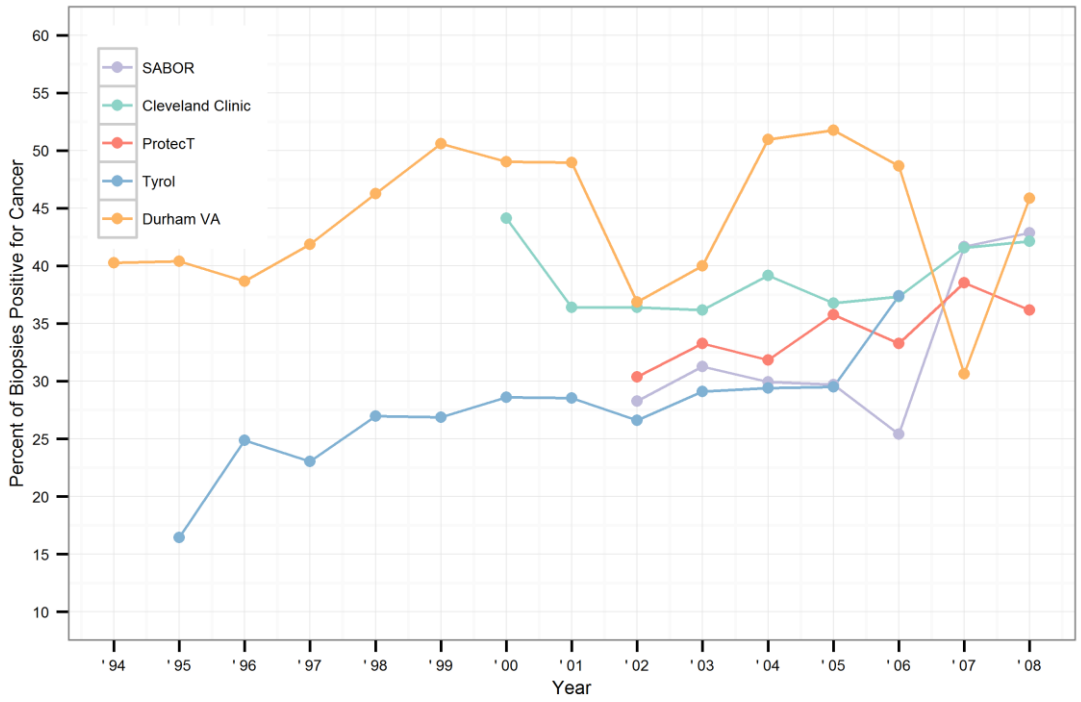
The five PBCG cohorts collected between 898 (SABOR) and 7260 (ProtecT) biopsies in the years 1994-2010 (Table 1). The two clinically referred cohorts, Cleveland Clinic and Durham VA, showed higher cancer rates, 39% and 46% respectively, than the three other, primarily screening, cohorts (27%-35%) (Figure 1). As expected, the PSA values were also higher in those two cohorts. Some biopsies in SABOR and almost half of the biopsies in Durham had missing

DRE. ProtecT did not collect DRE results at all. Family history was not present for Durham and Tyrol and the other three cohorts had missing values in 15%-40% of the cases. In the Austrian cohort, no information was available on the ethnicity but participants can be assumed to be primarily of Caucasian origin. Compared to the other cohorts, Durham VA had a remarkable representation of patients with African American origin (44%). In 20% of the cases patients had more than one biopsy. This fact was accounted for by the introduction of the risk factor prior biopsy. The data collection spanned timeframes between 8 years in ProtecT and 16 years in Durham VA. The yearly number of biopsies ranged from 73 to 1106.

Table 1. Biopsy characteristics from the five PBCG cohorts. Prostate specific antigen (PSA) measured in ng/ml. Risk factors used in the models for each cohort (≤ 15% missing); *SABOR & Cleveland Clinic:* PSA, age, DRE, race, prior biopsy; *ProtecT:* PSA, age, family history, race; *Tyrol:* PSA, age, prior biopsy, DRE; *Durham VA:* PSA, age, race, prior biopsy.

	SABOR N = 898	Cleveland Clinic N = 3257	ProtecT N = 7260	Tyrol N = 4749	Durham VA N = 2185
Age median (range)	64 (36, 89)	64 (50, 75)	63 (50, 72)	62 (50, 75)	64 (50, 75)
PSA* median (range)	3.2 (0.1, 49.8)	5.7 (0.2, 49.9)	4.3 (3.0, 49.7)	4.0 (0.2, 49.6)	5.1 (0.1, 49.5)
DRE result					
Normal	603 (67%)	3057 (94%)	0 (0%)	4392 (92%)	876 (40%)
Abnormal	234 (26%)	200 (6%)	0 (0%)	357 (8%)	251 (11%)
Unknown	61 (7%)	0 (0%)	7260 (100%)	0 (0%)	1058 (48%)
Family history	-	-	-	-	-
No	244 (27%)	1679 (52%)	5692 (78%)	0 (0%)	0 (0%)
Yes	295 (33%)	371 (11%)	453 (6%)	0 (0%)	0 (0%)
Unknown	359 (40%)	1207 (37%)	1115 (15%)	4749 (100%)	2185 (100%)
African origin					
No	794 (88%)	2799 (86%)	6878 (95%)	0 (0%)	1110 (51%)
Yes	104 (12%)	412 (13%)	31 (0%)	0 (0%)	963 (44%)
Unknown	0 (0%)	46 (1%)	351 (5%)	4749 (100%)	112 (5%)
Prior biopsy					
Yes	305 (34%)	1089 (33%)	0 (0%)	1417 (30%)	548 (25%)
No	593 (66%)	2168 (67%)	7260 (100%)	3332 (70%)	1637 (75%)
N cancer cases (%)	285 (32%)	1265 (39%)	2507 (35%)	1281 (27%)	963 (44%)

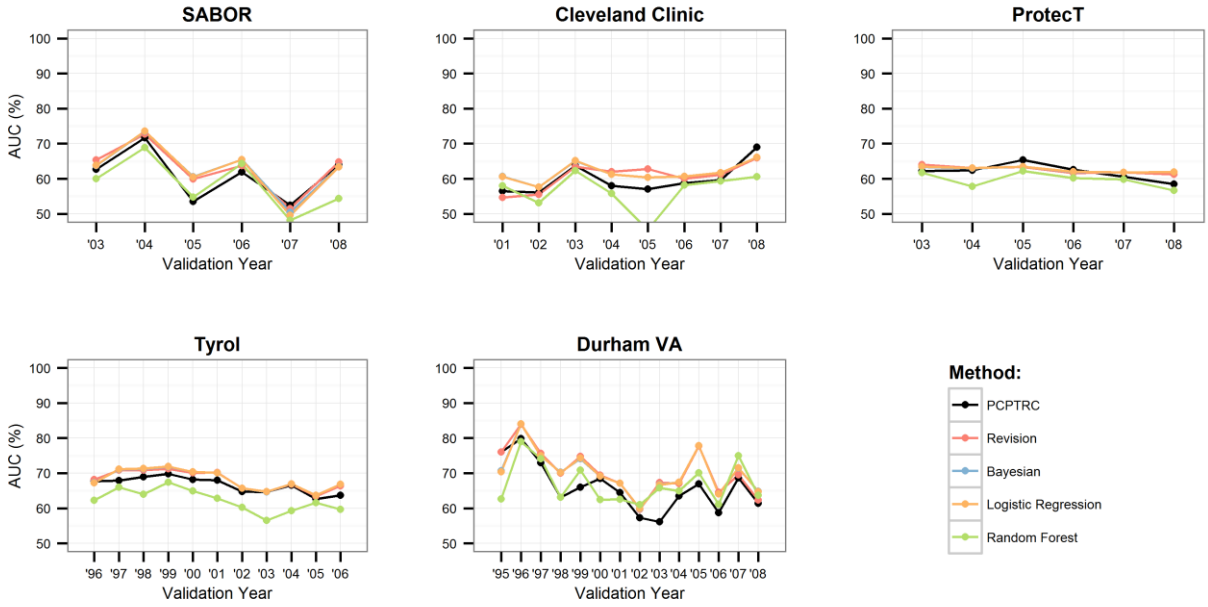
Figure 1. Yearly cancer rates for the five PBCG cohorts.



The six methods for dynamically updating a risk calculator were applied cumulatively to all years past in the cohort as a training set, with AUCs and HLSs evaluated on the next year as validation set. There were some large differences in validation performance for any given method. Focusing first on discrimination (Figure 2), the AUCs of methods evaluated on the SABOR data oscillated by up to 10 points across validation years and were almost at random performance (50%) in some years. Our expectation was that the prediction model would become more accurately trained to the cohort and the AUC would increase each year, but this was not the case for most of the cohorts. The logistic regression and the Bayesian updating exhibited almost identical performance throughout the cohorts. The AUCs of the revision method were comparable to the logistic regression method and the Bayesian approach, outperforming those in some years while not in others. Random forests were the worst performer in most cohorts, as they were consistently over-fitting the training data. Varying the tuning parameters did not help (data not shown). The static PCPTRC lagged behind the other 3 methods, all of which tailor to the institution, but was not statistically significantly inferior at the 95% level. The AUCs of the static PCPTRC and the recalibration were identical because recalibration is a monotonic transformation of the risk predictions. Differences in AUCs across cohorts were larger than differences among

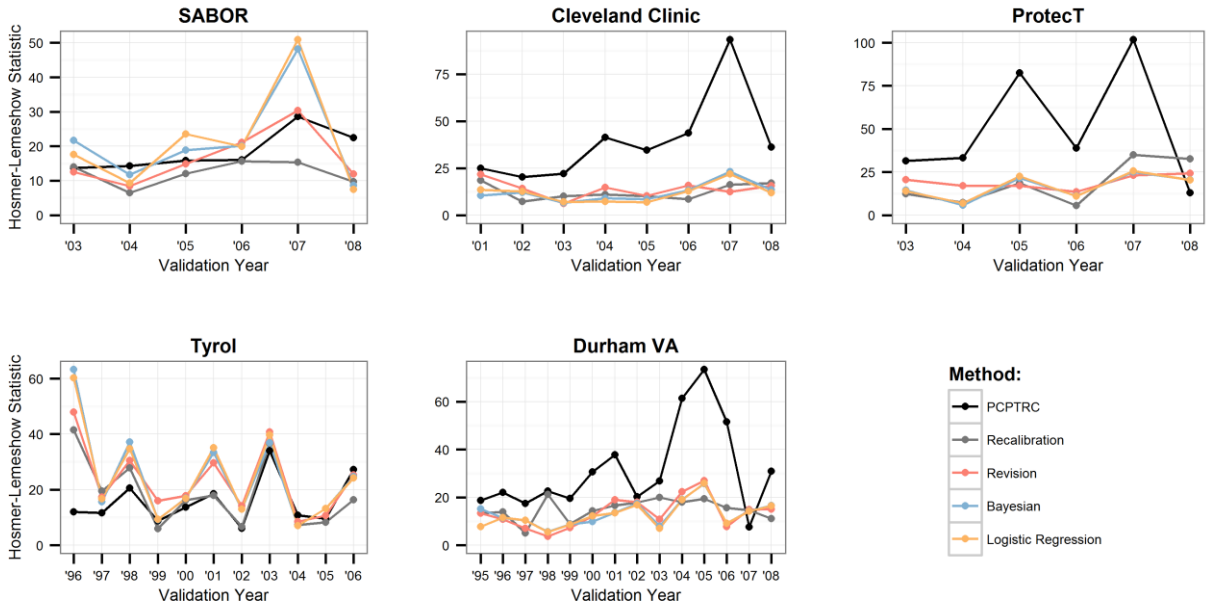
methods within any single cohort. The overall performance of all methods was worst on Cleveland Clinic and best on the Durham VA cohort.

Figure 2. AUCs using all past data as training and the next year as the validation year (x-axis). Higher values indicate better discrimination between cancer cases and control. The recalibration method gave identical results to the PCPTRC method.



In terms of calibration measured by the HLS, the random forest method performed so poorly that its values fell off the cohort graphs in Figure 3. The PCPTRC also performed statistically worse for the Cleveland Clinic, ProtecT and Durham VA cohorts. The PCPTRC performed better than the tailored approaches (logistic regression, Bayesian and revision) for some specific years in the SABOR cohort. In the early years of Tyrol, the static PCPTRC even outperformed recalibration. Typically these were the years where the cancer prevalence dramatically changed; see Figure 1. The tailored methods performed substantially worse after the abrupt change but adapted quickly so that performance was back to normal afterwards. By measuring calibration, closeness of expected to observed risks, on a squared loss scale rather than discrimination on a rank-based scale, the HLS was more sensitive for detecting changes in prediction that arise from sharp changes in prevalence or other characteristics.

Figure 3. Hosmer-Lemeshow test statistics using all past data as training and the next year as the validation year (x-axis). Lower values indicate better fit, closer agreement between observed and predicted risks. Random forests are omitted since their values were so high as to fall off the graph.



When fixed validation sets comprising the last 3 years were considered, the AUC increased with size of the training set for most cohorts, but the gain in AUC was small (Supplementary Appendix Figure 1). Revision, Bayesian and logistic regression performed equally well and consistently outperformed the static PCPTRC. In terms of calibration measured by the HLS, the static PCPTRC performed substantially worse than all other methods in the Cleveland Clinic, ProtecT and Durham VA cohorts (Supplementary Appendix Figure 2). HLS decreased for increasing size of the training set in the SABOR, Tyrol and Durham VA cohorts. In most cohorts, the recalibration method performed best.

4. DISCUSSION

In this report we have investigated alternative approaches for solving two problems facing contemporary clinical risk tools at once: the need for such tools to evolve over time to adapt to changes in clinical practice patterns, and the need for such tools to be tailored to accommodate local differences in patient-populations. What we have observed is that compared to static use of

a global prediction model for prostate cancer prediction, recalibration often improved calibration but had little impact on discrimination both in terms of temporal evaluation within an institution as well as across different institutions.

A large number of prior studies have evaluated the advantages of one-time revisions to update long-standing clinical models for new patients at the same institution or network of institutions. To give a few examples, in the context of predicting the risk of postoperative pain, the recalibration approach significantly improved calibration beyond that obtained by more complicated revision methods, but also had no effect on discrimination, as found here [24]. Update of a static coronary artery disease model in a contemporary network of 14 institutions increased calibration and maintained discrimination [26]. A one-time temporal recalibration of a mortality model following colorectal surgery improved calibration [28]. Recalibration of a pediatric mortality tool enhanced calibration in subgroups, which diminished discrimination [29]. Discrimination in all of these studies was measured by the AUC. Based on ranks of observations, the AUC has been notoriously proven to be difficult to budge [36]. The failure of the AUC to increase over time in all cohorts as training data accumulated over the years could be that the current risk factors collected for prostate cancer have reached their discrimination potential; it has often been noted that new markers are needed to substantially improve current risk prediction tools for prostate cancer [37].

There have been up to recently relatively fewer investigations of repeated temporal updates to existing clinical prediction models. Temporal quality control charts were used to monitor an intensive care unit score to monitor quality of care, with recalibration instigated when control measures exceeded bounds, and later extended to classification trees [30, 31]. This hinges on an interesting aspect not covered in this report – of diagnostic measures for assessing and only implementing recalibration when it is really needed. For a risk score predicting the mortality from cardiac surgery, repeated updates were performed to overcome the issue of calibration drift [38, 39]. Changes in the coefficients of the risk model were monitored for different temporal updating schemes, but performance measures for discrimination and calibration were not investigated. Recently in the informatics field approaches to transfer learning for adapting risk tools from one hospital to another have developed [32, 40]. These rely on global maximization of an objective function that sums over individual hospitals, allowing individual hospitals to collect different predictors. These were developed for the case of rare diseases, where the incidence is so low as to

demand synthesis of information across multiple hospitals, as well as to where no static risk calculator built on a single cohort is available.

Because of the large numbers of years and cohorts to make comparisons, we used only crude single number summaries, the AUC and HLS, to evaluate discrimination and calibration, respectively. In practice, more extensive detailed analyses should be implemented for investigating the performance of risk prediction tools. The AUC and HLS statistics summarize the more detailed and informative graphical displays, the receiver-operating characteristic curve (ROC) and calibration plot. The way risk predictions are used in practice is that a risk threshold is chosen, above which the patient is referred to further diagnostic testing or treatment. The ROC specifically reports the sensitivity (number of true cancer cases correctly referred) versus specificity (number of non-cancer cases that are correctly not referred) for every possible choice of a threshold. Evaluation of risk tools should rather be based on optimizing specificities/sensitivities for feasible thresholds than by optimizing the threshold-free AUC measure. Calibration plots are preferred over the HLS as they may show more detailed patterns of performance which can be summarized by Cox recalibration statistics, specifically an intercept reflecting calibration-in-the-large, and a calibration slope reflecting the overall strength of the predictors in the model [41]. There are many more methods for evaluating risk prediction tools, including the Brier score and net-benefit curves [23]. The Brier score is an integrated measure of discrimination and calibration. We repeated all analyses here using the Brier score as the outcome and arrived at nearly the same conclusions as to those based on the AUC (Figure 3 in the Supplementary Appendix). Net-benefit curves are geared towards finding optimal models for basing clinical decisions based on thresholds – they ultimately revert to combinations of sensitivity, specificity and prevalence, and so are also combined measures [42]. We have chosen the base measures of discrimination and calibration here because they represent the two pure most orthogonal components of model validation [43].

It has been established that shrinkage of regression coefficients or penalized regression may improve calibration of a risk-prediction tool by reducing the range of prediction values; see Chapter 13 of [23] for an overview. There are many possible options for performing shrinkage, but they all require an internal bootstrapping or cross-validation strategy to optimize tuning parameters. Shrinkage works similarly to the Bayesian method here, but estimates the amount of shrinkage from the data directly, instead of from a prior distribution. We tried a common method of shrinkage, the Lasso, which penalizes the regression by the L^1-norm of the parameter vector.

Discrimination performance was similar to the revision and Bayesian methods, and calibration performance was slightly worse than that of the recalibration method. These results were not surprising, because shrinkage has a bigger effect on calibration than on the AUC (a robust rank-based statistic that is difficult to move) and because our cohorts had only a small number of predictors (ranging from 3 to 5). However, a more thorough investigation of shrinkage methods is advisable for future applications.

Even though the validation sets used to evaluate the performance of the updating methods were taken from the same institution and chronologically in close proximity to the training sets, we still encountered large variations in cancer prevalence and other patient characteristics between training and test sets. In order to investigate if the sudden changes in cancer prevalence are accompanied by similar changes to the patient characteristics, we made plots for each cohort where we overlay cancer prevalence over time along with the prevalence of high covariate values (Supplementary Appendix Figure 4). However, we could not find clear associations between the spikes in cancer prevalence and changes of other characteristics. These unexpected variations in case-mix make it even harder to automatically evaluate model performance which is a key part of implementing unsupervised updating of risk scores in clinical practice. Recent efforts to create a framework for interpreting the results of external validation in the context of clinical prediction models may eventually lead to improved automation of the updating process [44].

In conclusion, a commonly available risk tool may provide adequate discrimination, but tailoring a risk model with institution-specific data may improve calibration. We recommend further implementation of updating methods to increase the accuracy of prediction models that are used in clinical practice.

REFERENCES

1.) Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation. 1998; 97(18): 1837-47.

2.) Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst. 1989; 81(24): 1879-86.

3.) Thompson IM, Ankerst DP, Chi C, Goodman PJ, Tangen CM, Lucia MS et al. Assessing prostate cancer risk: results from the Prostate Cancer Prevention Trial. J Natl Cancer Inst. 2006; 98(8): 529-34.

4.) Goff DC Jr, Lloyd-Jones DM, Bennett G, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. [published correction appears in J Am Coll Cardiol. 2014; 63(25 Pt B): 3026] J Am Coll Cardiol. 2014; 63(25 Pt B): 2935-59

5.) Ridker PM, Cook NR. Statins: new American guidelines for prevention of cardiovascular disease. Lancet 2013; 382(9907): 1762-5

6.) Vickers AJ, Fearn P, Kattan MW, Scardino PT. Why can't nomograms be more like Netflix? Urology. 2010; 75(3): 511-3.

7.) Thompson IM, Goodman PJ, Tangen CM, et al. The influence of finasteride on the development of prostate cancer. N Engl J Med. 2003; 349(3): 215-24.

8.) Thompson IM, Pauler DK, Goodman PJ, et al. Prevalence of prostate cancer among men with a prostate-specific antigen level < or = 4.0 ng per milliliter. N Engl J Med. 2004; 350(22): 2239-46.

9.) Parekh DJ, Ankerst DP, Higgins BA, et al. External validation of the Prostate Cancer Prevention Trial risk calculator in a screened population. Urology. 2006; 68(6): 1152-5.

10.) Eyre SJ, Ankerst DP, Wei JT, et al. Validation in a multiple urology practice cohort of the Prostate Cancer Prevention Trial calculator for predicting prostate cancer detection. J Urol. 2009; 182(6): 2653-8.

11.) Hernandez DJ, Han M, Humphreys EB, et al. Predicting the outcome of prostate biopsy: Comparison of a novel logistic regression-based model, the prostate cancer risk calculator, and prostate-specific antigen level alone. BJU Int. 2009; 103(5): 609-14.

12.) Cavadas V, Osorio L, Sabell F, et al. Prostate cancer prevention trial and European randomized study of screening for prostate cancer risk calculators: a performance comparison in a contemporary screened cohort. Eur Urol. 2010; 58(4): 551-8.

13.) Kaplan DJ, Boorjian SA, Ruth K, et al. Evaluation of the Prostate Cancer Prevention Trial Risk calculator in a high-risk screening population. BJU Int. 2010; 105(3): 334-7.

14.) Nam RK, Kattan MW, Chin JL, et al. Prospective multi-institutional study evaluating the performance of prostate cancer risk calculators. J Clin Oncol. 2011; 29(22): 2959-64.

15.) Trottier G, Roobol MJ, Lawrentschuk N, et al. Comparison of risk calculators from the Prostate Cancer Prevention Trial and the European Randomized Study of Screening for Prostate Cancer in a contemporary Canadian cohort. BJU Int. 2011; 108(8b): E237-44.

16.) Oliveira M, Marques V, Carvalho AP, et al. Head-to-head comparison of two online nomograms for prostate biopsy outcome prediction. BJU Int. 2011; 107(11): 1780-3.

17.) Zhu Y, Wang JY, Shen YJ, et al. External validation of the Prostate Cancer Prevention Trial and the European Randomized Study of Screening for Prostate risk calculators in a Chinese cohort. Asian J Androl. 2012; 14(5): 738-44.

18.) Ankerst DP, Boeck A, Freedland SJ, et al. Evaluating the PCPT risk calculator in ten international biopsy cohorts: results from the Prostate Biopsy Collaborative Group. World J Urol. 2012; 30(2): 181-7.

19.) Lee DH, Jung HB, Park JW, et al. Can Western based online prostate cancer risk calculators be used to predict prostate cancer after prostate biopsy for the Korean population? Yonsei Med J 2013; 54(3): 665-71.

20.) Ankerst DP, Boeck A, Freedland SJ, et al. Evaluating the Prostate Cancer Prevention Trial High Grade prostate cancer risk calculator in 10 international biopsy cohorts: results from the prostate biopsy collaborative group. World J Urol. 2014; 32(1): 185-91.

21.) Vickers AJ, Cronin AM, Roobol MJ, et al. The relationship between prostate-specific antigen and prostate cancer risk: The Prostate Biopsy Collaborative Group. Clin Cancer Res. 2010; 16(17): 4374-81.

22.) Ankerst DP, Till C, Boeck A, et al. The impact of prostate volume, number of biopsy cores and American Urological Association symptom score on the sensitivity of cancer

detection using the Prostate Cancer Prevention Trial risk calculator. J Urol. 2013; 190(1): 70-6.

23.) Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. New York: Springer; 2009.

24.) Janssen KJ, Moons KG, Kalkman CJ, et al. Updating methods improved the performance of a clinical prediction model in new patients. J Clin Epidemiol. 2008; 61(1): 76-86.

25.) Schuetz P, Koller M, Christ-Crain M, et al. Predicting mortality with pneumonia severity scores: importance of model recalibration to local settings. Epidemiol Infect. 2008; 136(12): 1628-37.

26.) Genders TS, Steyerberg EW, Alkadhi H, et al. A clinical prediction rule for the diagnosis of coronary artery disease: validation, updating, and extension. Eur Heart J. 2011; 32(11): 1316-30.

27.) Van Hoorde K, Vergouwe Y, Timmerman D, et al. Simple dichotomous updating methods improved the validity of polytomous prediction models. J Clin Epidemiol. 2013; 66(10): 1158-65.

28.) Kong CH, Guest GD, Stupart DA, et al. Recalibration and validation of a preoperative risk prediction model for mortality in major colorectal surgery. Dis Colon Rectum. 2013; 56(7):844-9.

29.) Visser IH, Hazelzet JA, Albers MJ, et al. Mortality prediction models for pediatric intensive care: comparison of overall and subgroup specific performance. Intensive Care Med. 2013; 39(5): 942-50.

30.) Minne L, Eslami S, de Keizer N, et al. Effect of changes over time in the performance of a customized SAPS-II model on the quality of care assessment. Intensive Care Med. 2012; 38(1):40-6.

31.) Minne L, Eslami S, de Keizer N, et al. Statistical process control for validating a classification tree model for predicting mortality — A novel approach towards temporal validation. J Biomed Inform. 2012; 45(1):37-44.

32.) Wiens J, Guttag J, Horvitz E. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. J Am Med Inform Assoc. 2014; 21(4):699-706.

33.) Ankerst DP, Hoefler J, Bock S, et al. Prostate Cancer Prevention Trial risk calculator 2.0 for the prediction of low- versus high-grade prostate cancer. Urology 2014; 83(6):1362-7.

34.) Pauler DK. The Schwarz criterion and related methods for Normal linear models. Biometrika 85:13-27, 1998.

35.) Breiman L. Random Forests. Machine Learning. 2001; 45(1): 5-32.

36.) Ware JH. The limitations of risk factors as prognostic tools. N Engl J Med. 2006; 355(25): 2615-7.

37.) Ankerst DP, Koniarski T, Liang Y, et al. Updating risk prediction tools: a case study in prostate cancer. Biom J. 2012; 54(1): 127-142.

38.) Hickey GL, Grant SW, Murphy GJ, et al. Dynamic trends in cardiac surgery: why the logistic EuroSCORE is no longer suitable for contemporary cardiac surgery and implications for future risk models. Eur J of Cardiothorac Surg. 2013; 43(6): 1146-1152.

39.) Hickey GL, Grant SW, Caiado C, et al. Dynamic prediction modeling approaches for cardiac surgery. Circ Cardiovasc Qual Outcomes. 2013; 6(6): 649-658.

40.) Lee G, Rubinfeld I, Syed Z. Adapting surgical models to individual hospitals using transfer learning. IEEE ICDM 2012 Workshop on Biological Data Mining and ist Applications in Healthcare (BioDM). 2012.

41.) Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. European heart journal. 2014; 35(29), 1925-31.

42.) Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006; 26(6): 565-574.

43.) Van Hoorde K, Van Huffel S, Timmerman D, et al. A spline-based tool to assess and visualize the calibration of multiclass risk predictions. J Biomed Inform. 2015; in press.

44.) Debray TP, Vergouwe Y, Koffijberg, H, et al. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol. 2015; 68(3): 279-89

FIGURE CAPTIONS

Figure 1. Yearly cancer rates for the five PBCG cohorts.

Figure 2. AUCs using all past data as training and the next year as the validation year (x-axis). Higher values indicate better discrimination between cancer cases and control. The recalibration method gave identical results to the PCPTRC method.

Figure 3. Hosmer-Lemeshow test statistics using all past data as training and the next year as the validation year (x-axis). Lower values indicate better fit, closer agreement between observed and predicted risks. Random forests are omitted since their values were so high as to fall off the graph.

Supplementary Appendix Figure 1. AUCs using all past data as training set and the last three years as the validation set. Latest year used for training corresponds to the x-axis. Higher values indicate better discrimination between cancer cases and controls.

Supplementary Appendix Figure 2. Hosmer-Lemeshow test statistics using all past data as training set and the last three years as the validation set. Latest year used for training corresponds to the x-axis. Lower values indicate better fit, closer agreement between observed and predicted risks. Random forests are omitted since their values were so high as to fall off the graph.

Supplementary Appendix Figure 3. Brier Score using all past data as training and the next year as the validation year (x-axis). Lower values indicate better fit.

Supplementary Appendix Figure 4. Changes of the patient characteristics over time for each cohort. The continuous variables PSA and age were dichotomized using the upper quartiles of PSA and age in the PBCG dataset.

Supplementary Appendix Figure 5. Patient characteristics from the year before until the year after a drastic change in cancer prevalence. The continuous variables PSA and age were dichotomized using the upper quartiles of PSA and age in the PBCG dataset.

Supplementary Appendix Figure 6. Patient Characterstics of the training and validation set for several years where the static PCPTRC outperformed the dynamically updated methods based on the Hosmer-Lemeshow statistic. The continuous variables PSA and age were dichotomized using the upper quartiles of PSA and age in the PBCG dataset.

Supplementary Appendix Figure 7. Yearly Hosmer-Lemeshow test statistics across validation years for the static PCPTRC and the recalibration method. The 95% confidence intervals are generated from 200 bootstrapped samples stratified by outcome.

Supplementary Appendix Figure 8. Hosmer-Lemeshow test statistics (HLS) for the static PCPTRC and the updating methods Recalibration and Revision in the cohorts Durham VA and SABOR. The updating process was repeated three times: all data (left panel), only patients without African American origin (middle panel) and just African American patients (right panel). SABOR had not enough African American participants to warrant a separate analysis. For Durham the updating and validation process was performed bi-annually to counteract the reduced sample size. Lesser values of the HLS correspond to better fit.

TABLE CAPTIONS

Table 1. Biopsy characteristics from the five PBCG cohorts. Prostate specific antigen (PSA) measured in ng/ml. Risk factors used in the models for each cohort (≤ 15% missing); *SABOR & Cleveland Clinic:* PSA, age, DRE, race, prior biopsy; *ProtecT:* PSA, age, family history, race; *Tyrol:* PSA, age, prior biopsy, DRE; *Durham VA:* PSA, age, race, prior biopsy.

Supplementary Appendix Table 1: PCPTRC formulas for prediction of cancer dependent on the available risk factors—this model was built from 6664 biopsies from the placebo arm of the Prostate Cancer Prevention Trial. Options M1 through M5 for the linear predictor M are allowed depending on what predictors are available.

Supplementary Appendix Table 2: Number of biopsies performed by year in the five PBCG cohorts.

Supplementary Appendix Table 3. Average AUC, range, and 95% confidence interval for the average AUC across all validation years for each cohort and method. The method with the highest AUC in each cohort is indicated in bold.

Supplementary Appendix Table 4. Average Hosmer-Lemeshow test statistic, range, and 95% confidence interval for the average AUC across all validation years for each cohort and method. The method with the highest AUC in each cohort is indicated in bold.

Supplementary Appendix Table 5. Parameters used for the `randomForest()` routine in the `R` package `randomForest`. We initially performed a Grid search to find a single set of parameters that would improve the method for all cohorts in all years. Since we couldn't find such a constellation, the default values of the package were used for the article.